LA-UR-02-3157

Approved for public release;
distribution is unlimited.

Title:	A BRIEF SURVEY OF STATISTICAL MODEL CALIBRATION IDEAS
Author(s):	Katherine Campbell
Submitted to:	Electronic to be published on the D-1 Statistical Computer Model Evaluation website

Los Alamos

NATIONAL LABORATORY

Form **836 (10/96)**

D R A F T

A Brief Survey of Statistical Model Calibration Ideas

Katherine Campbell
Statistical Sciences Group
Los Alamos National Laboratory

Context and nomenclature

We are here concerned about computerized models to be used in decision-making contexts of some sort. Examples are climate models and environmental models of other types (used to evaluate future CO_2 scenarios and or the effects of other perturbations of the biosphere), weapons models (used to certify stockpile performance), traffic models (used to evaluate the impact of proposed physical or policy changes to the transportation system), models of engineered systems (used to design new systems or improve existing ones), and so forth. For simplicity, and because we have so many other needs for the word "model", we will refer to such a computerized model as a "[computer] simulator" and to one run of the simulator as a "simulation".

Common features of such simulators are:

- A substantial theoretical and observational data base for the physical processes being simulated. This is true even if (as in the case of TRANSIMS, a traffic model) some of the dynamic modeling may be quite abstract, by comparison with the PDEs that are solved in the typical hydrologic or weapons simulator (which most tend to think of as more accurate mathematical representations of the physical dynamics.) The simulators that we will be concerned with are thus "physics-based".
- Very large numbers of input parameters, many or most of which are physically interpretable and either directly or indirectly measurable. Most often, for the large simulators that are typical of decision-making contexts, these physical parameters are parameters of submodels—for material strength, or the transmissivity of geologic materials, or passing behavior on dual-lane highways—that are constructed and estimated independently of the simulator. One of the major tasks in model evaluation is often to recover the uncertainties that were associated with these estimation processes. In the following discussion, we assume that this has been done, so that we have available empirically defensible distributions for these parameters (which we will refer to as prior or uncalibrated distributions.)
- Output on a very large number of potential observables. Only some of these are of interest to the decision maker, but there may be abundant and high-quality data on others. (If there were such data on the outputs of interest, then we would probably not be considering the use of computerized simulator as an aid to decision making!)

We restrict ourselves to "decision-making contexts" because we wish to assume that the intended use(s) of the computer simulator are well specified. Absent such specification, it becomes difficult to maintain necessary distinctions between

"calibration", "validation" and "prediction". There are certainly other uses for simulators. In particular, simulators of complex systems (i.e., systems that are too large and/or too nonlinear to be readily understood based on intuition or analytical calculations alone) are often built to provide insight for researchers in areas from immune systems to nuclear weapons. However, we are aiming at a more formal context, one in which the ultimate goal is to build a defensible statistical model for the uncertainty associated with simulator predictions.

A non-iterative paradigm for uncertainty quantification

A clean and philosophically unproblematic paradigm for the quantification of the total uncertainty associated with simulator predictions has the following steps. (In presenting this conceptual program, I do not intend to imply that its execution in practice is easy or even possible without a great many assumptions and simplifications.)

1. Prior uncertainty. Assemble prior information, including any application-specific (scenario-specific) information, on the simulator inputs and parameters. Represent uncertainty about these inputs and parameters by means of probability distributions. Generally the information being summarized is thought of as "model-independent", that is, based on theory, observation and/or expert judgment about the situation being modeled, but not on the specifics of the computer simulator. This is something of an idealization in some cases. For example, some of the inputs may be numerical parameters (such as mesh-stiffening parameters for Lagrangian calculations) or scale-dependent parameters (effective transmissivities in geologic flow and transport models) that clearly relate to the numerical algorithms and computational grids of the simulator. But even in these cases, the allowable ranges may sometimes be specified using expert judgment or scaling theory, more or less independently of the computer simulator.
2. Uncertainty propagation. Use the simulator to transform the joint input probability distribution into a joint distribution on the output variables. This transformation may be very difficult to compute in practice. It also may not be a deterministic transformation, if the model has stochastic components. The latter can add significantly to the computational difficulty of the problem but the basic idea is the same. Statistical emulators and simplified physics-based models can play a role in this computation.
3. Uncertainty quantification. Examine the relationship between the transformed distribution and observations on the output variables in order to characterize the distribution of $y_{computed}(x_{test}) - y_{nature}(x_{test})$, where $y_{computed}$ is a vector of simulated outputs and y_{nature} are the corresponding experimental quantities, over the range of scenarios x_{test} and outputs covered by in the validation data base. Such an uncertainty model should characterize at least the accuracy (bias) and precision (variance) of this distribution.
4. Prediction. Extrapolate to characterize the distribution $z_{computed}(x_{prediction}) - z_{nature}(x_{prediction})$ outside the range of the validation data base.

Note that not only is $x_{prediction}$ different from x_{test}, but also it is possible that z and y correspond to different simulator outputs. Nevertheless, some estimate of prediction variance, at least, is usually available. At a minimum, we have the propagated uncertainty (now for the z's) as in Step 2. If there were some observations on outputs of the type to be predicted in the validation data base (i.e., some overlap between the components of y and those of z), there may be additional information with which to construct an extrapolation model. However, formulating and estimating an explicit extrapolation model will be in general a very problem-specific task.

For this non-iterative paradigm, the prediction uncertainty resulting from Step 4 is likely to be very large (even if it is taken to include only the propagated uncertainty.) Sensitivity and importance analysis might tell us something about how to allocate resources to improve the information on the inputs in the most cost-effective way so that the output distributions from Step 2 may become more precise. Available observations on the model outputs are used in Step 3, and, if a non-trivial extrapolation model is constructed in Step 4, probably in that step as well. Estimates of validation and prediction errors can be cleanly partitioned into components due to uncertainty about model inputs, to numerical error, to measurement errors associated with data on outputs, and, in the case of prediction error, to additional uncertainties associated with the extrapolation model.

An iterative paradigm: model calibration

The non-iterative paradigm described above does not correspond to what is done in the real world. Nor does it really make optimal use of all of the data available for the simulator outputs, particularly when many of these data are related to outputs other than those for which prediction errors are ultimately required. Data on outputs of an integrated simulator are analogous to data from full system tests, and to use them only in validation mode is like using only component data to estimate the reliability of a system represented by a hierarchical fault tree model. The components interact in subsystems and in the full system in ways that cannot be fully predicted in advance. Therefore modern methods for uncertainty quantification need to address principled ways to use data on simulator outputs for calibration as well as validation.

Informally, such information has almost certainly been used during the construction of the simulator. For example, because an initial run of an Atlantic Ocean model produced currents in the wrong direction through the Florida Straits, local modifications of the representation of basin topography were needed, even though the topography was initially represented using natural and consistent rules for the relatively coarse grid (although at 0.1^o to 0.5^o, the resolution of this model was actually pushing the state of the art for ocean models.) These types of gross adjustments, frequently to numerical and effective or averaged parameters (see below; in the present example, gridded topography is an effective parameter) are regular features of model construction, and trouble almost no one.

Model calibration is a more formal version of the same process. From the statistical perspective of this review, calibration means updating the prior, simulator-independent information about the inputs (from Step 1 above) as a result of comparing the simulator

output with the available calibration data, and iterating back through Steps 1 and 2 before proceding with the characterization of the [calibrated] simulator uncertainty in Step 3. This sounds like an ordinary statistical regression problem, but to the extent that the simulator is "physics-based" and its parameters represent measurable quantities, such a statistical [re]estimation of its parameters appears harder to justify. Beyond asserting that the calibration data set should not be the same as the validation data set used in Step 3 (and there are typically many ways in which the available data on the various simulator outputs might be split into a calibration subset and a validation subset), what other principles are needed to guide this process?

It is very useful, first of all, to recognize that there are actually several types of input variables that must be prescribed before a computer simulator can be run.

- Scenario descriptors. These include forcing terms such as pollutant source descriptions, transportation network modifications, or gradients of magnetic quadrupoles for an accelerator model. Also included are specifications for the outputs to be computed: position, time, frequency, etc. As a rule, some of these scenario variables will be different in the prediction simulations from those used in the calibration/validation simulations. The defining presumption distinguishing these inputs from those more usually called parameters is that they are either known, or else they can be drawn from a user-specified distribution (a distribution representing the user's domain of interest or the unpredictability of the state of nature for a possible future scenario, rather than the user's uncertainty about constants of nature), for a given simulation.

- Physical parameters (constants of nature). These should, in theory, be constant across scenarios, because they represent underlying physical constants, although due to measurement problems their true values may be poorly known. The word "parameter" is used inclusively here. It may include indicator parameters for selecting among two or more alternative parameterizations of some physical process. It may include heterogeneous fields of distributed parameters. It might also include initial or boundary conditions if these are not variable across the domain of interest (i.e., not scenario dependent.)

  It is important to note that even physical parameters come in several flavors. Some can be interpreted as physical constants that are more or less directly measurable (the speed of light, or the density of a material at standard temperature and pressure), or are estimable as expected values over a population (the mean energy of a proton beam, or a rate constant for a chemical interaction.) Others are the result of fitting empirical models with several parameters to data. For example, the yield strength Y of a material increases with total plastic strain ε according to a convex form that is sometimes modeled with three parameters: $Y = Y_0 (1+\beta\varepsilon)^n$. Yet other physical parameters are averaged to represent a heterogeneous reality at the grid scale (topography in the ocean model example, or the effective permeability of a geologic material.) Such effective or conceptual parameters are not easily inferred from quantities that can be directly measured.

For example, not only averaging but also some kind of scaling may be required to get from measured values to effective values suitable for discretized simulators.

- Other simulator parameters. In general, for simulators of any complexity, there are additional parameters under the control of the modeler for which the specifications are incomplete. These include numerical parameters such as time steps and controls for adaptive mesh refinement or other numerical algorithms. They also may include some "semi-physical" parameters, sometimes with suggestive names like "artificial viscosity" or "friction coefficients", but in general these do not correspond to measurable quantities; rather they are numerical parameters required to make the numerical algorithms behave appropriately.

Once this wide range in the nature of the parameters of the simulator is recognized, it becomes apparent that there is quite a bit of leeway for calibration, even in a "physics-based" simulator. Inputs from both the second and third of the categories above are candidates for calibration, although some should certainly be more severely constrained by prior knowledge and available data than others.

Still, there remain conceptual difficulties with the whole idea of calibration using an integrated simulator. (The very idea of calibration has influential detractors, including Kleijnan 2001, who believes that at least for certain types of simulators, "calibration is considered bad practice".)

- For one thing, calibrated input distributions are no longer model-independent. If they relate to physical parameters, they should not, without many caveats, be supposed to update information about the corresponding physical quantities obtained from independent physical experiments. Rather, to the extent that such submodel parameters are updated by calibration, the updated distributions define the range that seems to "work" best when a given submodel is combined with others in an integrated simulator. So this is already quite different from the way we would interpret, for example, results from Bayesian updating of the parameters of a physics submodel as new data become available.
- Another very major conceptual problem in applying the Bayesian updating framework or another statistical estimation paradigm is that for complex simulations the model output typically deviates systematically, not randomly, from the observations, across all parameterizations within the prior distributions. In general these deviations are also larger than can be accomodated by the measurement error model. In short, the empirical "likelihood" of the result of a simulation is nill. This is, of course, an indication of model inadequacy, and may also suggest where model improvement might be sought. But the interesting question in the decision context is whether it is possible to estimate a "model discrepancy" term so that an existing simulator (which may represent many man-years of development effort) may nevertheless have value as an aid to decision making.

The Kennedy and O'Hagan framework

The statistical approach to model calibration that comes closest to recognizing all of the problems outlined above was initiated in the work of Kennedy and O'Hagan (2001; hereafter KOH). KOH divide the simulator inputs into two classes: what were called scenario descriptors above (x), assumed to be known or else sampled from a user-specified distribution, and all other model parameters (θ), which are assumed to be constant across scenarios but imperfectly known. They also estimate both a model discrepancy term and also, separately, experimental error. Thus their model for the available calibration data $y(x_{test})$ becomes

$$y(x_{test}) = S_Y(x_{test}, \theta) + \delta_Y(x_{test}) + \varepsilon_Y(x_{test}) \quad (1)$$

where $S_Y(x, \theta)$ is the output on the variables Y calculated by the simulator given inputs (x,θ), $\delta_Y(x)$ is the model discrepancy for these variables for scenario x, and $\varepsilon_Y(x)$ is the experimental error associated with the measurement of these variables for scenario x.

[KOH assume that the simulator is deterministic. However, they also discuss at considerable length the possibility that S_Y may be difficult to compute, and allow for the possibility of simultaneously estimating the parameters of a statistical emulator. A stochastic component for S_Y could be handled similarly. In addition, the KOH version of Eq. (1) includes a estimable scalar multiplier ρ for the simulator output $S_Y(x, \theta)$, apparently to accomodate the possibility that the simulator is systematically in error with respect to scale. We ignore these additional complications for the time being.]

KOH suggest modeling $\delta_Y(x)$ as a Gaussian process (i.e., one completely characterized by a mean function $E\delta_Y(x)$, often assumed to be zero, and a covariance function $Cov\{\delta_Y(x), \delta_Y(x')\}$) indexed by a subset of the scenario variables. With this specification, Eq. (1) resembles a statistical "mixed model" with "fixed effects" θ and "random effects" δ_Y, a relatively familiar statistical object. KOH take a Bayesian approach to estimating the statistical parameters associated with δ_Y and ε_Y together with the fixed effects θ.

In KOH examples, the scenario variables indexing the Gaussian process term $\delta_Y(x)$ take values in a Euclidean space (i.e., space and/or time) and $\delta_Y(x)$ is modeled with a translation-invariant covariance function. The calibration and prediction scenarios differ only with respect to these indexing variables, and the prediction outputs are the same as the calibration outputs (or at least a subset of them.) Thus for this special case calibration also provides a complete extrapolation model. The model for prediction is simply

$$z(x_{pred}) = y(x_{pred}) = S_Y(x_{pred}, \theta) + \delta_Y(x_{pred}) \quad (2)$$

where for θ we might use the mode of the posterior distribution. Uncertainty in this prediction is estimated from the uncertainty represented by this posterior distribution plus the uncertainty in extrapolation δ_Y from x_{test} to x_{pred}. If x_{pred} is available only as a user-

specified distribution (KOH call this case "risk analysis"), then that contributes a third component to the prediction uncertainty.

The situation will clearly be more complicated when x_{pred} differs from x_{test} in respects that are more difficult to quantify, or when the z's are different from the y's. However, provided there is some way of estimating a cross-covariance function $\mathrm{Cov}\{\delta_Y(x), \delta_Z(x')\}$, an extension might be possible. A non-conservative alternative would be to use $\delta_Z(x) = E\delta_Z(x) = 0$, so that the uncertainty in the prediction comes only from propagation of the updated input distributions.

As described by KOH, the calibration process simultaneously updates the prior distributions for θ and estimates the covariance function for the random effects term $\delta_Y(x)$. In the framework of the preceding section we might have identified these as separate steps (calibration, which updates θ, and validation, which estimates δ_Y). However, there are significant advantages to estimating θ and δ_Y simultaneously, in order to avoid "overtuning" the parameters θ. A subsequent validation-only step using additional validation data could be undertaken to improve the estimate of δ_Y (quite possibly this "improvement" using data not available to the calibration process would result in inflating its variance) without further updating θ.

Other approaches

The work of KOH is almost unique in admitting and explicitly estimating an empirical model-discrepancy term. Others, such as Finsterle and Persoff 1997, augment their models with additional terms to account for observed systematic errors, but these terms are explicitly associated with unmeasured deviations between experimental conditions and modeled conditions, e.g., leakage from the apparatus. The KOH model discrepancy term is simply an empirical description of model inadequancies from all, not necessarily identified, sources. The Kennedy and O'Hagan paper (together with some subsequent extensions, mentioned below) is certainly the most formal approach to model calibration to appear in the statistical literature, but other approaches with a statistical flavor do appear in other scientific literatures. These approaches fall into several groups, although the following classification is somewhat arbitrary and there is more similarity among them than may be obvious from the short descriptions offered below.

Parameter optimization Typical of optimization methods is the work of Cooley and Vecchia (1987, 1999), also implemented in software authored by Doherty et al. (1999). In these papers, optimization of and confidence intervals for the simulator parameters are based on nonlinear least squares regression calculations and extended to provide prediction intervals for functionals of the parameters (i.e., other model outputs). Typically the interval predictions are based on search over a region in which a weighted sum of squared residuals for the calibration data is less than some selected value above the minimum. If the model is capable of making predictions within measurement errors then the this upper bound might be justified by likelihood arguments, but frequently in practice the choice is *ad hoc*.

For nonlinear models with large numbers of parameters, these regions may be non-convex, even multiply connected, and difficult to explore. Therefore the associated computational issues are a major subject of research in applications. Approaches include the use of statistical emulators (again, Gaussian process models are a popular option, for example Cox et al. 2001) and genetic algorithms (Duan et al., 1992, and many since).

Monte Carlo methods These methods attempt to sample the distributions of the predictions of a calibrated simulator rather than providing any point optimization. Sometimes these approaches are explicitly Bayesian, based on statistical models for errors in the observations and the model (e.g., Romanowicz et al. 1994. See also the work of Poole and Raftery 2000, which employs Rubin's sampling-importance-resampling algorithm for a similar purpose.) In many cases, however, the *ad hoc* nature of the "likelihood" measure is acknowledged (as in most applications of the "Generalized Likelihood Uncertainty Estimation" method originating with Beven and Binley 1992.) A given simulation is weighted in the final estimation according to the "likelihood" that the calibration observations could have been produced by it. For some of these methods the weights come from a continuum; for others they are 1 or 0, pass or fail. The 0-1 methods are favored by those who are most interested in exploring a complex (non-convex, multiply connected) domain within the parameter space that appears to produce "behavioral" results (e.g., Spear et al. 1994.)

Monte Carlo methods are also employed by researchers who treat the problem explicitly as an ill-posed inverse problem. This approach seems to be most popular in oil reservoir problems (Oliver et al. 1997, Glimm et al, 2001), where it is applied to a field of distributed parameters, to which a Gausian process prior distribution is typically assigned.

System identification Some attempts have been made to cast the calibration problem as a data assimilation problem, using a formalism like the extended Kalman filter. These methods appear to have been most successful when more calibration data become available over time for a model, such as a catchment model for which the calibration data are stream gauge data corresponding to various storm events. Examples are found in papers by Stigler and Beck (1994), Banks (2001) and also (although not explicitly traced in this case to the data assimilation literature) Thiemann et al. (2001.) Here is Stigter and Beck's succinct description of the Kalman filter:

> "Quintessentially, the [Kalman] filter reconciles a prediction from the model with an observation of the system through a process of feedback, in which the account taken of the mismatch between theory and observation is modulated according to the balance of uncertainties attaching to these two elements of knowledge."

Although this sounds like a description of the model calibration problem, published applications of Kalman filtering and data assimilation look quite different from the problems considered above. Nevertheless, there are important similarities. In particular, KOH's reduction of the problem to a fairly standard statistical "mixed model" brings the problem much closer to the Kalman formulation. (The link between the estimation of random effects and the Kalman filter is well recognized; see for example Robinson 1991.)

Some more recent extensions of KOH work using Bayes linear methods look even more Kalman-like (Craig et al. 2001, Goldstein and Rougier preprint.)

An illustrative example

We conclude with a discussion of an example that illustrates many of the preceding remarks. A groundwater model for the Pajarito Plateau is being developed at the Los Alamos National Laboratory primarily to address problems associated with the transport of contaminants resulting from historical Laboratory operations to production wells supplying the town of Los Alamos. This model addresses only flow and transport in the saturated zone, several hundred meters below the Plateau on which most of the Laboratory and townsite facilities are located. The complex geology of the region is captured using a grid with vertical resolution of about 12 m, but the horizontal resolution is considerably larger (250 m.) Within the hydrosratigraphic units (about **??** in all) resolved at this scale, permeability and porosity are treated as constants, for which there are a handful of measurements (most at the laboratory scale.) Boundary conditions are generally flow- or no-flow boundaries, but on the west side of the model boundary flow estimates have been derived from a basin-scale regional model. The surface infiltration model is an elevation-dependent with a handful of independent parameters for which the prior information consists largely or entirely of informed judgement.

This is a simulator which has many of the properties mentioned above. It is a physics-based model, integrating many years of both qualitative and quantitative geological and hydrological studies of the Pajarito Plateau and Española Basin. It is not a simple computation. One run typically requires 20-40 minutes on a Sun workstation, and at least three runs are necessary to model three scenarios associated with a steady-state past (pre-development, up to about 1950), the transient conditions of the more recent past (the historical development phase, 1950 to 2000), and whatever future scenario is of interest (assuming steady-state extraction at 2000 rates, or some alternative future development scenario.) But at the same time, major simplifications have been made, including but not limited to the above-mentioned assumptions about the homogeneity of hydrostratigraphic units and the simplified infiltration model, and there are clearly some important but unobservable boundary conditions. The physical parameters of the model range from effective versions of measurable quantities such as permeability to the empirical parameterization of infiltration as a function of elevation. Scenario parameters include historical and projected pumping rates at various production wells.

Of interest in the decision context is the transport of contaminants from sources (generally to the west and south of Los Alamos) to environmentally accessible points (either production wells or the Rio Grande, which forms the southeast boundary of the modeled area.) However, the data available for calibrating and validating this model are almost exclusively related to static water levels (SWLs), some of which have been measured repeatedly over time for several decades. To date there have been few observations of contaminants in the groundwater, and when contaminants have been unambiguously observed it is unclear what their source was. So even if we restrict ourselves to predicting the transport of non-sorbing contaminants (a worst-case scenario, as most of the potential contaminants interact with and are retarded by the geological

materials in the aquifer and in the overlying vadose zone), the available data do not correspond to the outputs of greatest interest. The Los Alamos National Laboratory is undertaking a more systematic groundwater sampling program which may produce more information in the future. However, the model in general predicts flow times from source to observation point well in excess of the 60 years for which the Laboratory has been in existence, so failure to observe contamination except perhaps in wells very close to sources is what the model would predict for the foreseeable future.

One simulation, or a sequence of past-to-future simulations as described above, can produce an abundance of outputs, including [time-varying] SWLs at various points in the aquifer as well as flow paths. (The delineation of flow paths by "particle tracking" is one aspect of the computation that can be stochastic.) Experimentally-based "prior" ranges for the critical input parameters (permeabilities of the hydrogeologic units and the parameters of the infiltration model) result in very large ranges for most of these outputs. The box plots in Figure 1 summarize these ranges for static water levels under the pre-development scenario in 28 wells scattered across the Pajarito Plateau. (The red line shows the observed values.)

Calibration of this model has been attempted by two methods. The first is a parameter optimization approach using the PEST software (Doherty et al. 1999.) This approach reduces the prior ranges of some of the parameters by factors of five or more, based on search within a region where the weighted sum of square residuals between simulator outputs and calibration data exceeds its minimum value by not more than ?%. Figure 2 shows some of these results (the magenta "pre-calibration" and green "post-calibration" lines, representing 95% confidence intervals. However, the simulator was

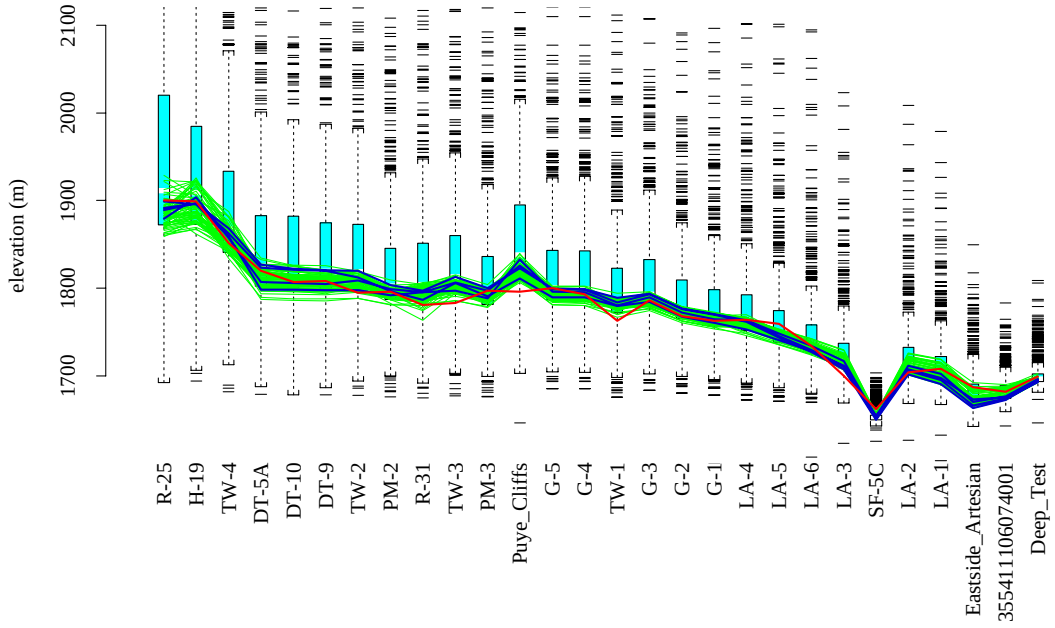

Figure 1. Predicted static water levels, pre-1950

modified between these two time-points, combining the two units labeled Tpf and Tpt, so these pairs are not comparable. Several of the magenta lines extend beyond the boundaries of this plot, as indicated by arrowheads.)

The second method is a Monte Carlo, GLUE-like method (using the earlier version of the simulator.) Figure 2 also shows parameter ranges for a large number or runs that passed some minimal convergence criteria (blue lines), while the box plots show ranges for 70 runs with the highest "likelihoods" (based on a somewhat different pair of criteria than the one used in parameter optimization.) The green and blue curves in Figure 1 also represent these 70 runs; the blue curves are the handful that also met the parameter-optimization criterion. Comparison with the observations (the red line in Figure 1) illustrates some of the systematic problems of the simulator. It is unable, for example, to match the observation in the Puye Cliffs well in any run that does a reasonable job overall; however, this well is far to the north of the area of interest and perhaps not of great concern. It lacks the heterogeneity needed to reproduce some patterns within closely-spaced groups of wells (such as LA-4, -5 and –6).

Both methods result in a wide spread of flow paths and times, as illustrated in Figure 3, which shows 70 (non-stochastic) flow paths from one potential source to one of two production wells or to the Rio Grande, color coded by travel time. (The red paths correspond to the runs that met the parameter optimization criterion.) These show that calibration by either method leaves a great deal of residual uncertainty about the prediction of the outputs of interest. Nevertheless some valuable insight has been gained. It is apparent from Figure 3 that surveillance wells that fail to penetrate the aquifer to a depth of several hundred meters is likely to fail to intercept contamination on its way to one of the two main production wells. The other feature of this particular set of data is that predicted travel times from this particular source to the accessible environment

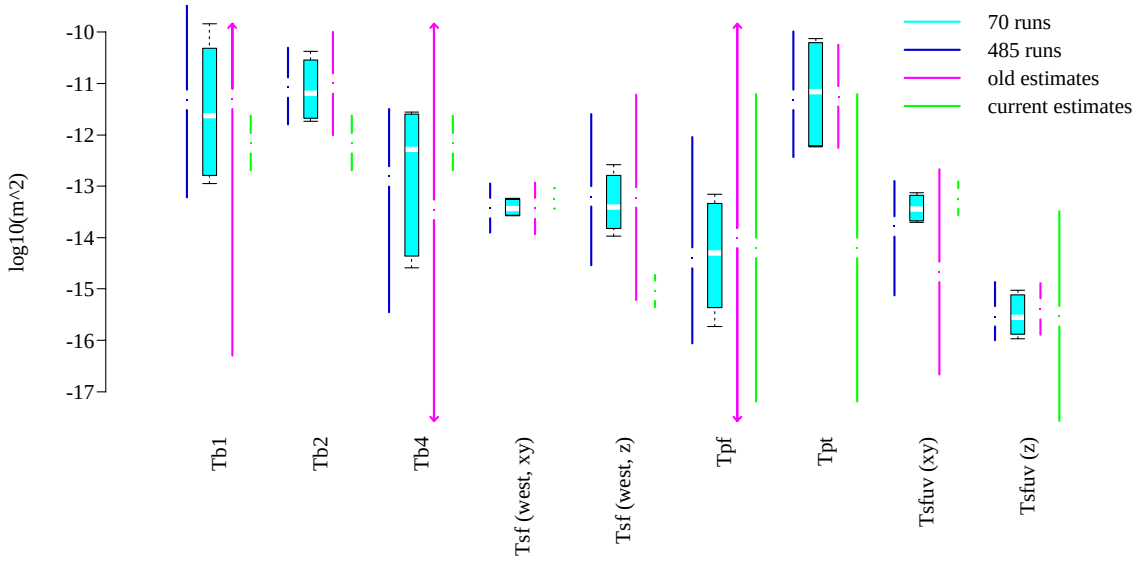

Figure 2. Ranges for selected calibration parameters

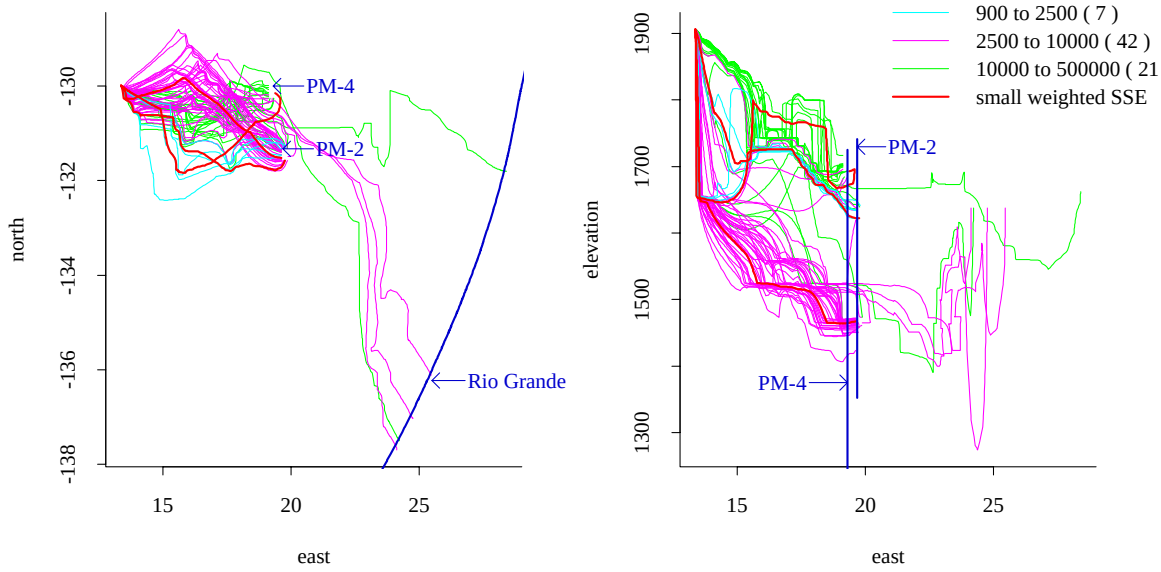

Figure 3. Flow paths from one potential source

generally exceed 1000 years, but of course there are other potential sources that are closer to production wells. So despite the inadequacies of the model, which continues to evolve, it already has some value in terms of integrating many sources of information about the groundwater system beneath the Pajarito Plateau. Observed SWLs are one type of information which essentially can be integrated only using some form of calibration.

References

Banks, H.T. (2001). Remarks on uncertainty assessment and management in modeling and computation. Mathematical and Computer Modelling **33**, 39-47.

Beven, K. and Binley, A. (1992). The future of distributed models: model calibration and uncertainty prediction. Hydrological Processes **6**, 279-298.

Cox, D.D., Park, J.-S. and Singer, C.E. (2001). A statistical method for tuning a computer code to a data base. Computational Statistics & Data Analysis **37**, 77-92.

Craig, P.S., Goldstein, M., Rougier, J.C. and Seheult, A.H. (2001). Bayesian forecasting for complex systems using computer simulators. J. Amer. Stat. Assoc. **96**, 717-729.

Doherty, J., Brebber, L. and Whyte, P. (1999). PEST: Model Independent Parameter Estimation. Watermark Computing, Third Edition. Brisbane, Australia.

Duan, Q.Y., Sorooshshian, S., and Gupta, V. (1992). Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resources Research **28**, 1015-1031.

Finsterle, S. and Persoff, P. (1997). Determining permeability of tight rock samples using inverse modeling. Water Resources Research **33**, 1803-1811.

Glimm, J., Hou, S., Lee, Y, Sharp, D. and Ye, K. (2000). Prediction of oil production with confidence intervals. Los Alamos Report LAUR-00-5583.

Goldstein, M. and Rougier, J. (preprint). Calibrated Bayesian forecasting using large computer simulators.

Kennedy, M.C. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion.) J. Royal Statistical Society (Series B) **63**, 425-464.

Kleijnan, J.P.C. (2001). Comments on M.C. Kennedy and A. O'Hagan's "Bayesian calibration of computer models". J. Royal Statistical Society (Series B) **63**.

Oliver, D.S., Cunha, L.B. and Reynolds, A.C. (1997). Markov chain Monte Carlo methods for conditioning a permeability field to pressure data. Mathematical Geology **29**, 61-91.

Poole, D. and Raftery, A.E. (2000). Inference for deterministic simulation models: the Bayesian melding approach. J. Amer. Stat. Assoc. **95**, 1244-1255.

Robinson, G.L. (1991) That BLUP is a Good Thing: the estimation of random effects. Statistical Science **6**, 15-51.

Romanowicz, R., Beven, K. and Tawn, J.A. (1994). Evaluation of predictive uncertainty in nonlinear hydrological models using a Bayesian approach. In *Statistics for the Environment 2: Water Related Issues* (V. Barnett and K.F. Turkman, eds.), John Wiley & Sons Ltd., 297-317.

Stigter, J.D. and Beck, M.B. (1994). A new approach to the identification of model structure. Environmetrics **5**, 315-333.

Spear, R.C., Grieb, T.M. and Shang, N. (1994). Parameter uncertainty and interaction in complex environmental models. Water Resources Research **30**, 3159-3169.

Thiemann, M., Trosset, M., Gupta, H. and Soroosian, S. (2001). Bayesian recursive parameter estimation for hydrologic models. Water Resources Research **37**, 2521-2535.